11. Logic of Hypothesis Testing

A. Introduction
B. Significance Testing
C. Type I and Type II Errors
D. One- and Two-Tailed Tests
E. Interpreting Significant Results
F. Interpreting Non-Significant Results
G. Steps in Hypothesis Testing
H. Significance Testing and Confidence Intervals
I. Misconceptions
J. Exercises

When interpreting an experimental finding, a natural question arises as to whether the finding could have occurred by chance. Hypothesis testing is a statistical procedure for testing whether chance is a plausible explanation of an experimental finding. Misconceptions about hypothesis testing are common among practitioners as well as students. To help prevent these misconceptions, this chapter goes into more detail about the logic of hypothesis testing than is typical for an introductory-level text.

Introduction

by David M. Lane

Prerequisites

- Chapter 5: Binomial Distribution

Learning Objectives

1. Describe the logic by which it can be concluded that someone can distinguish between two things
2. State whether random assignment ensures that all uncontrolled sources of variation will be equal
3. Define precisely what the probability is that is computed to reach the conclusion that a difference is not due to chance
4. Distinguish between the probability of an event and the probability of a state of the world
5. Define "null hypothesis"
6. Be able to determine the null hypothesis from a description of an experiment
7. Define "alternative hypothesis"

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from the binomial distribution and the binomial distribution calculator shows it to be 0.0106. This is a pretty low probability, and therefore

someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for normal-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in Chapter 12, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

> It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial distribution, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he

cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and therefore are difficult to apply in some contexts.

The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average}$$

or as

$$\mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers

hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The alternative hypothesis is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{obese} = \mu_{average}$$

is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average}$$
$$\mu_{obese} > \mu_{average}.$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two populations means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960) showed how it is justified to draw a conclusion about the direction of the difference.

Significance Testing

by David M. Lane

Prerequisites

- Chapter 5: Binomial Distribution
- Chapter 11: Introduction to Hypothesis Testing

Learning Objectives

1. Describe how a probability value is used to cast doubt on the null hypothesis
2. Define "statistically significant"
3. Distinguish between statistical significance and practical significance
4. Distinguish between two approaches to significance testing

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α level or simply α. It is also called the *significance level*.

When the null hypothesis is rejected, the effect is said to be *statistically significant*. For example, in the Physicians Reactions case study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

> Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher) a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis. If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

Type I and II Errors

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing

Learning Objectives

1. Define Type I and Type II errors
2. Interpret significant and non-significant differences
3. Explain why the null hypothesis should not be accepted when the effect is not significant

In the Physicians' Reactions case study, the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05 then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α level or simply α. It is also called the significance level. As discussed in the introduction to hypothesis testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the α level: the lower the α level the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The probability of correctly rejecting a false null hypothesis equals $1 - \beta$ and is called power. Power is covered in detail in Chapter 13.

One- and Two-Tailed Tests

by David M. Lane

Prerequisites

- Chapter 6: Binomial Distribution
- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance

Learning Objectives

1. Define one- and two-tailed tests
2. State the difference between one- and two-tailed hypotheses
3. State when it is valid to use a one-tailed test

In the James Bond case study, Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure 1 shows a graph of the binomial. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a "one-tailed probability."

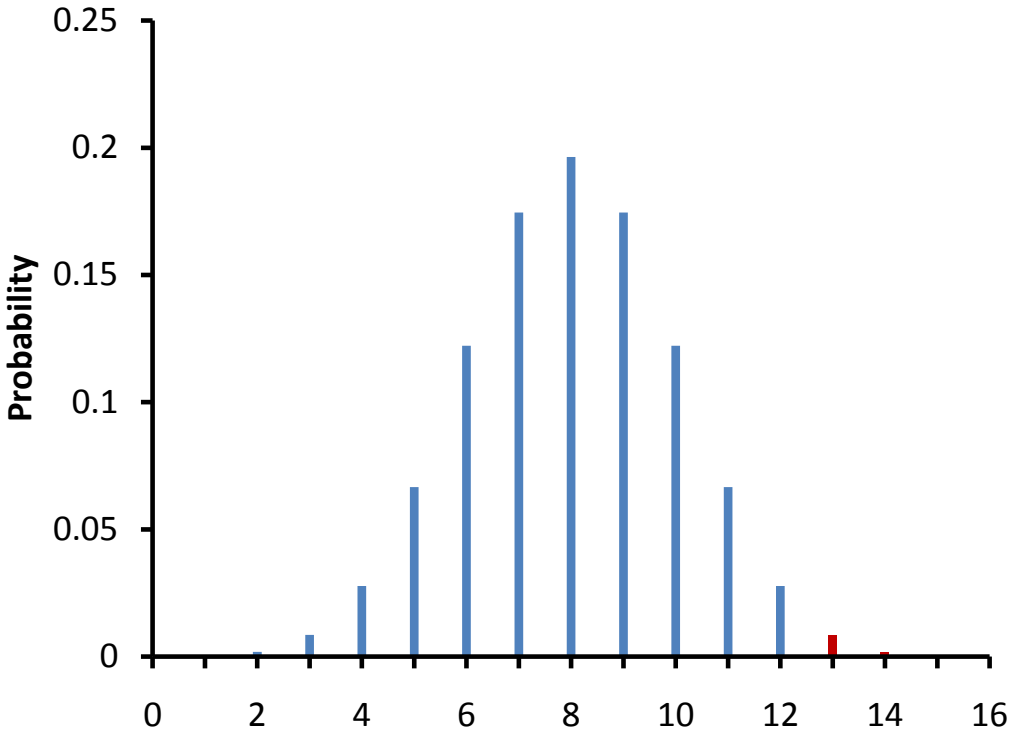


Figure 1. The binomial distribution. The upper (right-hand) tail is red.

A slightly different question can be asked of the data: "What is the probability of getting a result as extreme or more extreme than the one observed"? Since the chance expectation is 8/16, a result of 3/13 is equally as extreme as 13/16. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously. Therefore, $p = 0.0212$. A probability calculated in both tails of a distribution is called a two-tailed probability (see Figure 2).

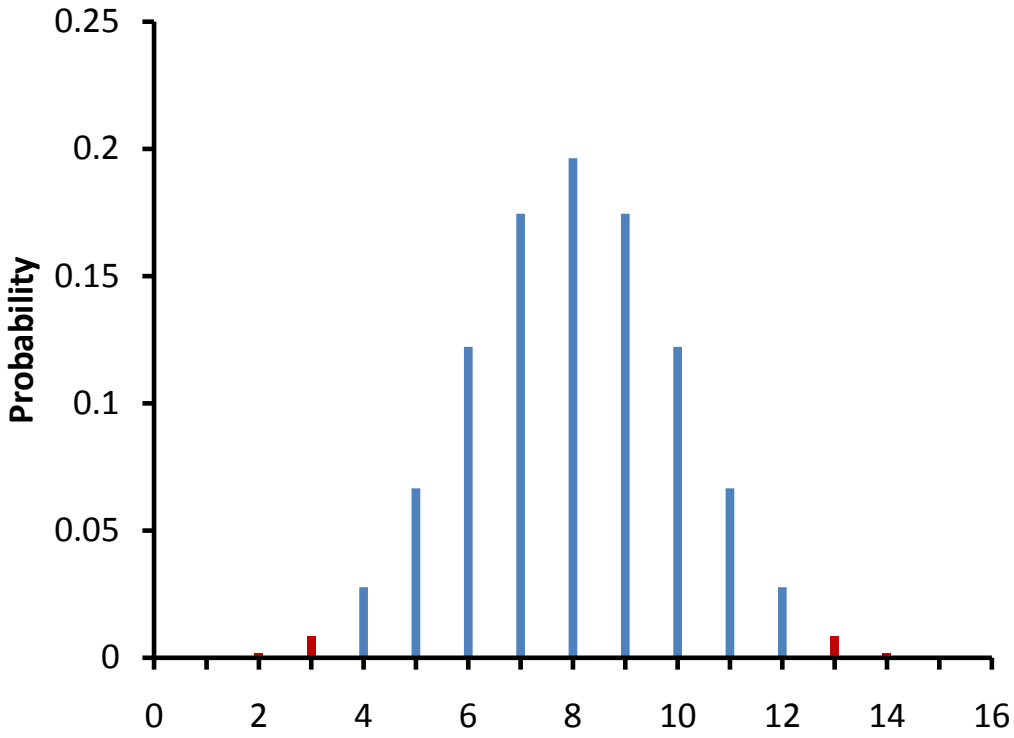


Figure 2. The binomial distribution. Both tails are red.

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond's performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is $\pi = 0.5$. By contrast, the null hypothesis for the one-tailed test is $\pi \leq 0.5$. Accordingly, we reject the two-tailed hypothesis if the sample proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected only if the sample proportion is much greater

than 0.5. The alternative hypothesis in the two-tailed test is $\pi \neq 0.5$. In the one-tailed test it is $\pi > 0.5$.

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

Interpreting Significant Results

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance, Type I and II Errors
- Chapter 11: One and Two-Tailed Tests

Learning Objectives

1. Discuss whether rejection of the null hypothesis should be an all-or-none proposition
2. State the usefulness of a significance test when it is extremely likely that the null hypothesis of no difference is false even before doing the experiment

When a probability value is below the α level, the effect is *statistically significant* and the null hypothesis is rejected. However, not all statistically significant effects should be treated the same way. For example, you should have less confidence that the null hypothesis is false if $p = 0.049$ than $p = 0.003$. Thus, rejecting the null hypothesis is not an all-or-none proposition.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. Consider the one-tailed test in the James Bond case study: Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred and the question is whether he is better than chance on this task. The null hypothesis for this one-tailed test is that $\pi \leq 0.5$ where π is the probability of being correct on any given trial. If this null hypothesis is rejected, then the alternative hypothesis that $\pi > 0.5$ is accepted. If π is greater than 0.5, then Mr. Bond is better than chance on this task.

Now consider the two-tailed test used in the Physicians' Reactions case study. The null hypothesis is:

$$\mu_{obese} = \mu_{average}.$$

If this null hypothesis is rejected, then there are two alternatives:

$$\mu_{obese} < \mu_{average}$$
$$\mu_{obese} > \mu_{average}.$$

Naturally, the direction of the sample means determines which alternative is adopted. If the sample mean for the obese patients is significantly lower than the sample mean for the average-weight patients, then one should conclude that the population mean for the obese patients is lower than the sample mean for the average-weight patients.

There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established. This point is also made in the section on the relationship between confidence intervals and significance tests.

Optional

Some textbooks have incorrectly stated that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Instead, they say that all one can conclude is that the population means differ. The validity of concluding the direction of the effect is clear if you note that a two-tailed test at the 0.05 level is equivalent to two separate one-tailed tests each at the 0.025 level. The two null hypotheses are then

$$\mu_{obese} \geq \mu_{average}$$
$$\mu_{obese} \leq \mu_{average}.$$

If the former of these is rejected, then the conclusion is that the population mean for obese patients is lower than that for average-weight patients. If the latter is rejected, then the conclusion is that the population mean for obese patients is higher than that for average-weight patients. See Kaiser (1960).

Interpreting Non-Significant Results

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing
- Chapter 11: Type I and II Errors

Learning Objectives

1. State what it means to accept the null hypothesis
2. Explain why the null hypothesis should not be accepted
3. Describe how a non-significant result can increase confidence that the null hypothesis is false
4. Discuss the problems of affirming a negative conclusion

When a significance test results in a high probability value, it means that the data provide little or no evidence that the null hypothesis is false. However, the high probability value is not evidence that the null hypothesis is true. The problem is that it is impossible to distinguish a null effect from a very small effect. For example, in the James Bond Case Study, suppose Mr. Bond is, in fact, just barely better than chance at judging whether a martini was shaken or stirred. Assume he has a 0.51 probability of being correct on a given trial ($\pi = 0.51$). Let's say Experimenter Jones (who did not know $\pi = 0.51$) tested Mr. Bond and found he was correct 49 times out of 100 tries. How would the significance test come out? The experimenter's significance test would be based on the assumption that Mr. Bond has a 0.50 probability of being correct on each trial ($\pi = 0.50$). Given this assumption, the probability of his being correct 49 or more times out of 100 is 0.62. This means that the probability value is 0.62, a value very much higher than the conventional significance level of 0.05. This result, therefore, does not give even a hint that the null hypothesis is false. However, we know (but Experimenter Jones does not) that $\pi = 0.51$ and not 0.50 and therefore that the null hypothesis is false. So, if Experimenter Jones had concluded that the null hypothesis was true based on the statistical analysis, he or she would have been mistaken. Concluding that the null hypothesis is true is called *accepting the null hypothesis*. To do so is a serious error.

Do not accept the null hypothesis when you do not reject it.

So how should the non-significant result be interpreted? The experimenter should report that there is no credible evidence Mr. Bond can tell whether a martini was shaken or stirred, but that there is no proof that he cannot. It is generally impossible to prove a negative. What if I claimed to have been Socrates in an earlier life? Since I have no evidence for this claim, I would have great difficulty convincing anyone that it is true. However, no one would be able to prove definitively that I was not.

Often a non-significant finding increases one's confidence that the null hypothesis is false. Consider the following hypothetical example. A researcher develops a treatment for anxiety that he or she believes is better than the traditional treatment. A study is conducted to test the relative effectiveness of the two treatments: 20 subjects are randomly divided into two groups of 10. One group receives the new treatment and the other receives the traditional treatment. The mean anxiety level is lower for those receiving the new treatment than for those receiving the traditional treatment. However, the difference is not significant. The statistical analysis shows that a difference as large or larger than the one obtained in the experiment would occur 11% of the time even if there were no true difference between the treatments. In other words, the probability value is 0.11. A naive researcher would interpret this finding as evidence that the new treatment is no more effective than the traditional treatment. However, the sophisticated researcher, although disappointed that the effect was not significant, would be encouraged that the new treatment led to less anxiety than the traditional treatment. The data support the thesis that the new treatment is better than the traditional one even though the effect is not statistically significant. This researcher should have more confidence that the new treatment is better than he or she had before the experiment was conducted. However, the support is weak and the data are inconclusive. What should the researcher do? A reasonable course of action would be to do the experiment again. Let's say the researcher repeated the experiment and again found the new treatment was better than the traditional treatment. However, once again the effect was not significant and this time the probability value was 0.07. The naive researcher would think that two out of two experiments failed to find significance and therefore the new treatment is unlikely to be better than the traditional treatment. The sophisticated researcher would note that two out of two times the new treatment was better than the traditional treatment. Moreover, two experiments each providing weak support that the new treatment is better, when

taken together, can provide strong support. Using a method for combining probabilities, it can be determined that combining the probability values of 0.11 and 0.07 results in a probability value of 0.045. Therefore, these two non-significant findings taken together result in a significant finding.

Although there is never a statistical basis for concluding that an effect is exactly zero, a statistical analysis can demonstrate that an effect is most likely small. This is done by computing a confidence interval. If all effect sizes in the interval are small, then it can be concluded that the effect is small. For example, suppose an experiment tested the effectiveness of a treatment for insomnia. Assume that the mean time to fall asleep was 2 minutes shorter for those receiving the treatment than for those in the control group and that this difference was not significant. If the 95% confidence interval ranged from -4 to 8 minutes, then the researcher would be justified in concluding that the benefit is eight minutes or less. However, the researcher would not be justified in concluding the null hypothesis is true, or even that it was supported.

Steps in Hypothesis Testing

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

Learning Objectives

1. Be able to state the null hypothesis for both one-tailed and two-tailed tests
2. Differentiate between a significance level and a probability level
3. State the four steps involved in significance testing

1. The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu_1 - \mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is $\mu_2 - \mu_1 \geq 0$. This is equivalent to $\mu_1 \leq \mu_2$.
2. The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.
3. The third step is to compute the probability value (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
4. Finally, compare the probability value with the α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

Significance Testing and Confidence Intervals

by David M. Lane

Prerequisites

- Chapter 10: Confidence Intervals Introduction
- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing

Learning Objectives

1. Determine from a confidence interval whether a test is significant
2. Explain why a confidence interval makes clear that one should not accept the null hypothesis

There is a close relationship between confidence intervals and significance tests. Specifically, if a statistic is significantly different from 0 at the 0.05 level then the 95% confidence interval will not contain 0. All values in the confidence interval are plausible values for the parameter whereas values outside the interval are rejected as plausible values for the parameter. In the Physicians' Reactions case study, the 95% confidence interval for the difference between means extends from 2.00 to 11.26. Therefore, any value lower than 2.00 or higher than 11.26 is rejected as a plausible value for the population difference between means. Since zero is lower than 2.00, it is rejected as a plausible value and a test of the null hypothesis that there is no difference between means is significant. It turns out that the p value is 0.0057. There is a similar relationship between the 99% confidence interval and significance at the 0.01 level.

Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative). Therefore, a significant finding allows the researcher to specify the direction of the effect. There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established because the values in the confidence interval are either all positive or all negative.

If the 95% confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then the effect will not be significant at the 0.05 level. Looking at non-significant effects in terms of confidence intervals makes clear why the null hypothesis should not be accepted when it is not rejected: Every value in the confidence interval is a plausible value of the parameter. Since zero is in the interval, it cannot be rejected. However, there is an infinite number of other values in the interval (assuming continuous measurement), and none of them can be rejected either.

Misconceptions

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

Learning Objectives

1. State why the probability value is not the probability the null hypothesis is false
2. Explain why a low probability value does not necessarily mean there is a large effect
3. Explain why a non-significant outcome does not mean the null hypothesis is probably true

Misconceptions about significance testing are common. This section lists three important ones.

1. Misconception: The probability value is the probability that the null hypothesis is false.
Proper interpretation: The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.

2. Misconception: A low probability value indicates a large effect.
Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.

3. Misconception: A non-significant outcome means that the null hypothesis is probably true.
Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 11: Interpreting Non-Significant Results

Research in March, 2012 reported here found evidence for the existence of the Higgs Boson particle. However, the evidence for the existence of the particle was not statistically significant.

What do you think?

Did the researchers conclude that their investigation had been a failure or did they conclude they have evidence of the particle, just not strong enough evidence to draw a confident conclusion?

> One of the investigators stated, "We see some tantalizing evidence but not significant enough to make a stronger statement." Therefore, they were encouraged by the result. In a subsequent study, the evidence was significant.

References

Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, 67, 160-167

Exercises

Prerequisites

- All material presented in the Logic of Hypothesis Testing chapter

1. An experiment is conducted to test the claim that James Bond can taste the difference between a Martini that is shaken and one that is stirred. What is the null hypothesis?

2. The following explanation is incorrect. What three words should be added to make it correct?

   The probability value is the probability of obtaining a statistic as different (add three words here) from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming that the null hypothesis is true.

3. Why do experimenters test hypotheses they think are false?

4. State the null hypothesis for:

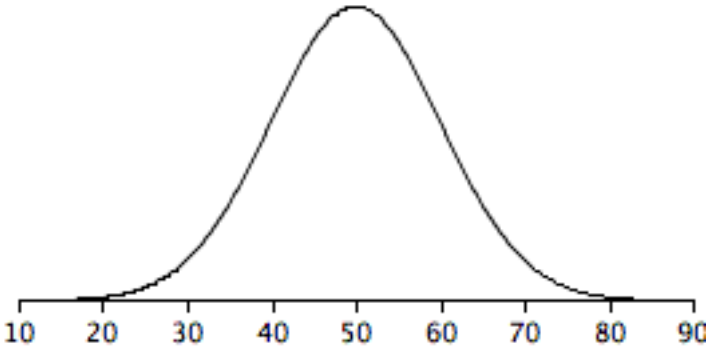
   a. An experiment testing whether echinacea decreases the length of colds.

   b. A correlational study on the relationship between brain size and intelligence.

   c. An investigation of whether a self-proclaimed psychic can predict the outcome of a coin flip.

   d. A study comparing a drug with a placebo on the amount of pain relief. (A one-tailed test was used.)

5. Assume the null hypothesis is that $\mu = 50$ and that the graph shown below is the sampling distribution of the mean (M). Would a sample value of M= 60 be significant in a two-tailed test at the .05 level? Roughly what value of M would be needed to be significant?



6. A researcher develops a new theory that predicts that vegetarians will have more of a particular vitamin in their blood than non-vegetarians. An experiment is conducted and vegetarians do have more of the vitamin, but the difference is not significant. The probability value is 0.13. Should the experimenter's confidence in the theory increase, decrease, or stay the same?

7. A researcher hypothesizes that the lowering in cholesterol associated with weight loss is really due to exercise. To test this, the researcher carefully controls for exercise while com- paring the cholesterol levels of a group of subjects who lose weight by dieting with a control group that does not diet. The difference between groups in cholesterol is not significant. Can the researcher claim that weight loss has no effect?

8. A significance test is performed and $p = .20$. Why can't the experimenter claim that the probability that the null hypothesis is true is .20?

9. For a drug to be approved by the FDA, the drug must be shown to be safe and effective. If the drug is significantly more effective than a placebo, then the drug is deemed effective. What do you know about the effectiveness of a drug once it has been approved by the FDA (assuming that there has not been a Type I error)?

10. When is it valid to use a one-tailed test? What is the advantage of a one-tailed test? Give an example of a null hypothesis that would be tested by a one-tailed test.

11. Distinguish between probability value and significance level.

12. Suppose a study was conducted on the effectiveness of a class on "How to take tests." The SAT scores of an experimental group and a control group were

compared. (There were 100 subjects in each group.) The mean score of the experimental group was 503 and the mean score of the control group was 499. The difference between means was found to be significant, p = .037. What do you conclude about the effectiveness of the class?

13. Is it more conservative to use an alpha level of .01 or an alpha level of .05? Would beta be higher for an alpha of .05 or for an alpha of .01?

14. Why is "Ho: "M1 = M2" not a proper null hypothesis?

15. An experimenter expects an effect to come out in a certain direction. Is this sufficient basis for using a one-tailed test? Why or why not?

16. How do the Type I and Type II error rates of one-tailed and two-tailed tests differ?

17. A two-tailed probability is .03. What is the one-tailed probability if the effect were in the specified direction? What would it be if the effect were in the other direction?

18. You choose an alpha level of .01 and then analyze your data.

    a. What is the probability that you will make a Type I error given that the null hypothesis is true?

    b. What is the probability that you will make a Type I error given that the null hypothesis is false?

19. Why doesn't it make sense to test the hypothesis that the sample mean is 42?

20. True/false: It is easier to reject the null hypothesis if the researcher uses a smaller alpha (α) level.

21. True/false: You are more likely to make a Type I error when using a small sample than when using a large sample.

22. True/false: You accept the alternative hypothesis when you reject the null hypothesis.

23. True/false: You do not accept the null hypothesis when you fail to reject it.

24. True/false: A researcher risks making a Type I error any time the null hypothesis is rejected.